Published in final edited form as:

J Appl Stat. 2012 December 1; 39(1): 2441–2452. doi:10.1080/02664763.2012.712954.

On Inference for Kendall's τ within a Longitudinal Data Setting

Yan Ma

Hospital for Special Surgery, Department of Public Health, Weill Medical College of Cornell University, New York, NY 10021

Summary

Kendall's τ is a non-parametric measure of correlation based on ranks and is used in a wide range of research disciplines. Although methods are available for making inference about Kendall's τ , none has been extended to modeling multiple Kendall's τ s arising in longitudinal data analysis. Compounding this problem is the pervasive issue of missing data in such study designs. In this paper, we develop a novel approach to provide inference about Kendall's τ within a longitudinal study setting under both complete and missing data. The proposed approach is illustrated with simulated data and applied to an HIV prevention study.

Keywords

HIV prevention; Inverse probability weighting; Kendall's τ ; Missing at random; U-statistics

1 Introduction

As a measure of association, Kendall's τ (Kendall, 1938) is utilized to measure rank correlation between the two quantities x and y of a bivariate random variable, $z = (x, y)^T$, in a broad spectrum of research fields including climatology (Huntington *et al.* 2004), economics (Pirttila and Uusitalo, 2010), mental health (Drake *et al.* 1989), oncology (Sipahi *et al.*, 2010), pediatrics (Mankuta *et al.* 2010), pharmacogenomics (Anfosso *et al.* 2006), and physics (Paulus *et al.* 2001), among others. Unlike the popular Pearson's correlation coefficient, relying on the product of moments to capture the relationship between two quantities, Kendall's τ is a non-parametric measure of correlation computed based on the ranks of x and y , and therefore is distribution-free. Further, it does not assume a linear relationship between x and y , and thus can be applied to both continuous and discrete bivariate random variables.

Since the introduction of the rank correlation coefficient, many related statistical methods have been developed to identify τ 's asymptotic properties (Hoeffding, 1948; Wood, 1970); to study the higher-order moments of τ (Snow, 1962); to examine publication bias in meta-analysis (Begg and Mazumdar, 1994); to generalize τ for testing independence against positive quadrant dependence (Kocher and Gupta, 1987) and for testing against serial dependence in a time series context (Ferguson *et al.* 2000). In addition, Kendall's τ has been extended to assessing association in genome data (Kang and Sen, 2008; Zhang *et al.* 2010).

Longitudinal studies that collect data repeatedly over time have become a driving force in modern-day research. However, none of the existing methods provides inference for dependent Kendall's τ s arising from longitudinal studies with missing data. As a result, the lack of methods to systematically address this problem has hampered the progress of research in the aforementioned fields. For example, in HIV prevention research, frequencies of risky behaviors such as unprotected vaginal sex reported over monthly or quarterly intervals are often compared to those from daily diary to help assess the degree of cognitive

bias in such proxy outcomes (Morrison-Beedy *et al.* 2006). With missing data in the study resulting from loss to follow-up, it is challenging to make inferences about the dependent Kendall's τ s originating from repeated assessments of the extent of concordance between the two recall measurements. In this situation, the most commonly utilized approach is to estimate or compare the correlated Kendall's τ s based on the subsample consisting of those subjects with complete data, which not only reduces power but may yield biased estimates as well.

Since the seminal work by Hoeffding (Hoeffding, 1948), U-statistics have been used to estimate a wide range of statistics such as order statistics (e.g., Mann-Whitney-Wilcoxon statistic, Lehmann 1951), the receiver operating characteristics (ROC) curve (Delong *et al.* 1988), and functions of higher order moments (e.g., concordance correlation coefficient, King *et al.* 2007, Ma *et al.* 2010; Kappa coefficient, Ma *et al.* 2008). In this paper, a new estimate of Kendall's τ rank correlation coefficient is developed based on U-statistics within a longitudinal study setting under both complete and missing data in Section 2. The approach is illustrated with simulated data in Section 3 and the HIV prevention study data in Section 4. Some remarks are given in Section 5.

2 Modeling Kendall's τ in a Longitudinal Data Setting

2.1 Notations of Kendall's τ in a Longitudinal Data Setting

We first consider an i.i.d sample with bivariate ordinal outcomes $\mathbf{z}_i = (x_i; y_i)^\top$ in a cross-sectional data setting ($i = 1, 2, \dots, n$). Kendall's τ is defined as

$$\tau = \frac{p_c - p_d}{\sqrt{(1 - p_x)(1 - p_y)}}, \quad (1)$$

where $p_c = \Pr((x_i - x_j)(y_i - y_j) > 0)$, $p_d = \Pr((x_i - x_j)(y_i - y_j) < 0)$, $p_x = \Pr(x_i = x_j)$, $p_y = \Pr(y_i = y_j)$ representing the probability of concordant (p_c) and discordant (p_d) pairs, and the probability of ties in x (p_x) and y (p_y), respectively. The value of Kendall's τ lies in $[-1, 1]$, with 1 implying perfect concordance, -1 perfect discordance, and 0 an independent relationship between the paired ordinal outcomes x and y . When $\mathbf{z}_i$ are continuous outcomes, Kendall's τ can be simplified to

$$\tau = 2p_c - 1, \quad (2)$$

which refers to Kendall's τ_a while the τ in (1) is called Kendall's τ_b (Kendall and Gibbons, 1990). The difference between τ_a and τ_b is that the latter controls for tied ranks in discrete data, whereas the former does not. In this paper, we focus on τ_b by considering τ_a a special case of τ_b .

We next consider Kendall's τ_b in a longitudinal setting with bivariate outcome variable over T assessments.

Let

$$\mathbf{z}_{it} = (x_{it}, y_{it})^\top, \mathbf{z}_i = (\mathbf{z}_{i1}^\top, \mathbf{z}_{i2}^\top, \dots, \mathbf{z}_{iT}^\top)^\top, \mathbf{z} = (\mathbf{z}_1^\top, \mathbf{z}_2^\top, \dots, \mathbf{z}_n^\top)^\top,$$

where $\mathbf{z}_{it}$ represents the bivariate outcome of subject i at time t , $i = 1, 2, \dots, n$; $t = 1, 2, \dots, T$.

Let

$$p_{ct} = \Pr((x_{it} - x_{jt})(y_{it} - y_{jt}) > 0), \quad p_{dt} = \Pr((x_{it} - x_{jt})(y_{it} - y_{jt}) < 0),$$

$$p_{xt} = \Pr(x_{it} = x_{jt}), \quad p_{yt} = \Pr(y_{it} = y_{jt}), \quad \mathbf{p}_c = (p_{c1}, p_{c2}, \dots, p_{cT})^\top,$$

$$\mathbf{p}_d = (p_{d1}, p_{d2}, \dots, p_{dT})^\top, \mathbf{p}_x = (p_{x1}, p_{x2}, \dots, p_{xT})^\top, \mathbf{p}_y = (p_{y1}, p_{y2}, \dots, p_{yT})^\top.$$

Kendall's τ_b in a longitudinal data setting is defined as

$$\tau_{bt} = \frac{p_{ct} - p_{dt}}{\sqrt{(1 - p_{xt})(1 - p_{yt})}}, \quad \boldsymbol{\tau}_b = (\tau_{b1}, \tau_{b2}, \dots, \tau_{bT})^\top,$$

where τ_{bt} is Kendall's τ_b at time t , $1 \leq t \leq T$.

2.2 Inference for Kendall's τ with Complete Data

In this section, we propose a new estimate of Kendall's τ and derive its asymptotic property within a longitudinal data setting under complete data.

Let

$$\eta_{1t} = p_{ct} - p_{dt}, \eta_{2t} = p_{xt}, \eta_{3t} = p_{yt}, \quad (3)$$

$$\boldsymbol{\eta}_t = (\eta_{1t}, \eta_{2t}, \eta_{3t})^\top, \boldsymbol{\eta} = (\boldsymbol{\eta}_1^\top, \boldsymbol{\eta}_2^\top, \dots, \boldsymbol{\eta}_T^\top)^\top,$$

$$f(\boldsymbol{\eta}_t) = \frac{\eta_{1t}}{\sqrt{(1 - \eta_{2t})(1 - \eta_{3t})}}, \quad \mathbf{f}(\boldsymbol{\eta}) = (f(\boldsymbol{\eta}_1), f(\boldsymbol{\eta}_2), \dots, f(\boldsymbol{\eta}_T))^\top.$$

Then, $\boldsymbol{\tau}_b = \mathbf{f}(\boldsymbol{\eta})$. Also, let

$$h_{1ijt} = I_{\{(x_{it} - x_{jt})(y_{it} - y_{jt}) > 0\}} - I_{\{(x_{it} - x_{jt})(y_{it} - y_{jt}) < 0\}}, \quad (4)$$

$$h_{2ijt} = I_{\{x_{it} - x_{jt}\}}, \quad h_{3ijt} = I_{\{y_{it} - y_{jt}\}},$$

$$\mathbf{h}_{ijt} = (h_{1ijt}, h_{2ijt}, h_{3ijt})^\top, \mathbf{h}_{ij} = (\mathbf{h}_{ij1}^\top, \mathbf{h}_{ij2}^\top, \dots, \mathbf{h}_{ijT}^\top)^\top, (i, j) \in C_2^n, 1 \leq t \leq T,$$

where $C_2^n = \{(i, j); 1 \leq i < j \leq n\}$ denotes the set of all distinct combinations of pairs (i, j) from the integer set $\{1, 2, \dots, n\}$. It is readily checked that $\boldsymbol{\eta} = E(\mathbf{h}_{ij})$. Further,

$\widehat{\boldsymbol{\eta}} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} \mathbf{h}_{ij}$ is an one-sample, two-argument, and vector-valued U-statistic. By

applying the theory of multivariate U-statistics (Hoeffding, 1948), $\hat{\tau}_b = f(\hat{\eta})$ is an unbiased, consistent, and asymptotically normal estimate of τ_b .

Of note, the proposed method can also be applied to Kendall's τ_a in (2) for continuous outcomes. Similarly, let $\tau_{at} = 2p_{ct} - 1$ denote Kendall's τ_a at time t in a longitudinal data setting, $1 \leq t \leq T$. With a kernel function $h_{ijt} = I_{\{(x_{it}-x_{jt})(y_{it}-y_{jt}) > 0\}}$, the τ_{at} can be estimated by

a U-statistic $\hat{\tau}_{at} = 2\hat{p}_{ct} - 1$, where $\hat{p}_{ct} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} h_{ijt}$.

2.3 Inference for Kendall's τ with Missing Data

Missing data is inevitable in longitudinal studies, and therefore choosing an appropriate approach for handling this issue is essential. In this section, we generalize the U-statistic based estimate of Kendall's τ for two popular missing data mechanisms—missing completely at random (MCAR) and missing at random (MAR).

To this end, define a vector of binary variables for indicating missing (or rather observed) response as follows:

$$r_{ixt} = \begin{cases} 1 & \text{if } x_{it} \text{ is observed} \\ 0 & \text{if } x_{it} \text{ is unobserved} \end{cases}, \quad r_{iyt} = \begin{cases} 1 & \text{if } y_{it} \text{ is observed} \\ 0 & \text{if } y_{it} \text{ is unobserved} \end{cases},$$

$$\mathbf{r}_{it} = (r_{ixt}, r_{iyt})^\top, \mathbf{r}_i = (\mathbf{r}_{i1}^\top, \mathbf{r}_{i2}^\top, \dots, \mathbf{r}_{iT}^\top)^\top, 1 \leq t \leq T.$$

Let

$$\Delta_{ixt} = \Pr[r_{ixt} = 1 | z_i], \Delta_{iyt} = \Pr[r_{iyt} = 1 | z_i], \Delta_{ixyt} = \Pr[r_{ixt}r_{iyt} = 1 | z_i],$$

$$\Delta_{it} = (\Delta_{ixt}, \Delta_{iyt}, \Delta_{ixyt})^\top, \Delta_i = (\Delta_{i1}^\top, \dots, \Delta_{iT}^\top)^\top, 1 \leq i \leq n.$$

We define our estimate as $\hat{\tau}_b = f(\hat{\eta})$ with $\hat{\eta}$ and $\hat{\tau}_b$ in a longitudinal study data setting given by:

$$\hat{\eta}_{1t} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} \frac{r_{ixt}r_{iyt}r_{jxt}r_{jyt}}{\Delta_{ixyt}\Delta_{jxyt}} h_{1ijt}, \hat{\eta}_{2t} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} \frac{r_{ixt}r_{jxt}}{\Delta_{ixt}\Delta_{jxt}} h_{2ijt} \quad (5)$$

$$\hat{\eta}_{3t} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} \frac{r_{iyt}r_{jyt}}{\Delta_{iyt}\Delta_{jyt}} h_{3ijt},$$

$$\hat{\eta}_t = (\hat{\eta}_{1t}, \hat{\eta}_{2t}, \hat{\eta}_{3t})^\top, \hat{\eta} = (\hat{\eta}_1^\top, \hat{\eta}_2^\top, \dots, \hat{\eta}_T^\top)^\top,$$

$$\widehat{\tau}_{bt} = f(\widehat{\eta}_t) = \frac{\widehat{\eta}_{1t}}{\sqrt{(1 - \widehat{\eta}_{2t})(1 - \widehat{\eta}_{3t})}}, \quad \widehat{\tau}_b = (\widehat{\tau}_{b1}, \widehat{\tau}_{b2}, \dots, \widehat{\tau}_{bT})^\top, \quad 1 \leq t \leq T.$$

Note that although h_{ijt} in (4) is not defined if one of the x_{it} , x_{jt} , y_{it} and y_{jt} is missing, $\widehat{\tau}_t$ above is well-defined, since h_{ijt} can be assigned any value in such cases without affecting $\widehat{\tau}_t$. The estimate $\widehat{\eta}$ in (5) may be viewed as a generalization of the classic *inverse probability weighted* (IPW) estimate used in non-parametric analysis to a U-statistics setting (e.g. Robins *et al.* 1995). In the absence of missing data, $\widehat{\tau}_t$ reduce to the estimates in the complete-data case and thus $\widehat{\tau}_b = f(\widehat{\eta})$ is a generalization of the estimate of τ_b under complete data.

By assuming τ_t is known we summarize the asymptotic property of $\widehat{\tau}_b$ in a theorem below, with a proof sketched in the supplementary materials.

Theorem 1.

(a) $\widehat{\tau}_b$ is consistent and asymptotically normal,

$$\widehat{\tau}_b \rightarrow_p \tau_b, \quad \sqrt{n}(\widehat{\tau}_b - \tau_b) \rightarrow_d N\left(\mathbf{0}, \sum_{\tau_b} = \frac{\partial^\top}{\partial \eta} f(\eta) \sum_{\eta} \frac{\partial}{\partial \eta} f(\eta)\right),$$

where $\rightarrow_p$ and $\rightarrow_d$ denote convergence in probability and in distribution, respectively, and Σ_{η} the covariance matrix of $\widehat{\eta}$.

(b) The asymptotic variance Σ_{τ_b} can be obtained by the Delta method. A consistent estimate of the covariance matrix $\Sigma_{\eta}(s, t) \equiv \text{Cov}(\eta_s, \eta_t)$, $1 \leq s, t \leq T$, can be obtained by

$$\widehat{\text{cov}}(\eta_s, \eta_t) = \widehat{E}\left(\mathbf{v}_{ijs} \mathbf{v}_{ikt}^\top\right), \quad (i, j, k) \in C_3^n, \quad (6)$$

where each component in the matrix $\widehat{\text{cov}}(\eta_s, \eta_t)$ can be expressed as an one-sample, three-argument U-statistic and $\mathbf{v}_{ijt}$ is defined as

$$\mathbf{v}_{ijt} = (v_{1ijt}, v_{2ijt}, v_{3ijt})^\top,$$

$$v_{1ijt} = \frac{r_{ixt} r_{iyt} r_{jxt} r_{jyt}}{\Delta_{ixyt} \Delta_{jxyt}} (h_{1ijt} - \eta_{1t}), \quad v_{2ijt} = \frac{r_{ixt} r_{jxt}}{\Delta_{ixt} \Delta_{jxt}} (h_{2ijt} - \eta_{2t}), \quad v_{3ijt} = \frac{r_{iyt} r_{jyt}}{\Delta_{iyt} \Delta_{jyt}} (h_{3ijt} - \eta_{3t}).$$

Theorem 1 is readily applied to test linear contrasts of the form:

$$H_0: K\tau_b = \mathbf{0}, \quad \text{vs.} \quad H_a: K\tau_b \neq \mathbf{0}, \quad (7)$$

where K is some $p \times T$ full rank matrix of known constants. For testing the hypothesis in (7), we can use the quadratic statistic, $Q_n^2 = n \widehat{\tau}_b^\top K^\top (K \widehat{\Sigma}_{\tau_b} K^\top)^{-1} K \widehat{\tau}_b$, which has an asymptotic central χ_2^2 distribution with p degrees of freedom under H_0 . For small samples, Q_n^2 often yields inflated type I errors, since it completely ignores the variability in the estimated $\widehat{\Sigma}_{\tau_b}$ (Guo *et al.*, 2005). A popular alternative for correcting this upward bias is the Hotelling's T-

square statistic, $T_n^2 = \frac{n-p}{p(n-1)} Q_n^2$, which follows approximately a central $F_{p, n-p}$ distribution with p (numerator) and $n-p$ (denominator) degrees of freedom under H_0 (e.g. Seber, 1984).

2.4 Estimation of the Weight Function

In most applications, Δ_j are unknown and must be estimated. Under MCAR, the r_i are independent of the z_i . Consequently, Δ_{ixt} , Δ_{iyt} and Δ_{ixyt} are all functionally independent of z_i and are readily estimated by the respective sample moments:

$$\widehat{\Delta}_{xt} = \frac{1}{n} \sum_{i=1}^n r_{ixt}, \widehat{\Delta}_{yt} = \frac{1}{n} \sum_{i=1}^n r_{iyt}, \text{ and } \widehat{\Delta}_{xyt} = \frac{1}{n} \sum_{i=1}^n r_{ixt} r_{iyt} \quad (1 \leq t \leq T).$$

When Δ_j becomes dependent on z_j , it is necessary to model Δ_j as a function of z_j . However, it is almost infeasible to model Δ_j as it may include too many missing data patterns without imposing some additional assumptions on the relationship between the occurrence of missing data and outcomes (Little and Rubin, 1987). In addition, the difficulty of modeling Δ_j is magnified by the increased number of missing data patterns when z_j is a bivariate outcome. In longitudinal trials missing data often occur as the result of subject dropout due to deteriorated/improved health and other related conditions, exhibiting the so-called monotone missing data pattern (MMDP). In such cases, missing data are predicted by observed responses and MAR arises naturally as a plausible model for predicting missing data. Further, the structured patterns under MMDP make it possible to model such a dependence in most studies.

In our study we focus on MAR mechanism and MMDP within a bivariate outcome setting. For each subject i , let $t_{x_i}^*$ and $t_{y_i}^*$ denote the last time point when x_i and y_i are observed, respectively. Let $l_i \equiv |t_{x_i}^* - t_{y_i}^*|$ denote the absolute difference or lag time between the last observed x_{it} and y_{it} and let $L = \max \{l_i; 1 \leq i \leq n\}$.

When $L = 0$, x_{it} and y_{it} are observed or missing as a pair so that $r_{ixt} = r_{iyt} = r_{it}$, and therefore $\Delta_{ixt} = \Delta_{iyt} = \Delta_{ixyt}$. Let $\tilde{z}_{it} = (z_{i1}^\top, \dots, z_{i(t-1)}^\top)^\top$ denote the history of observed response pairs prior to time t ($2 \leq t \leq T$). Under MMDP, if z_{it} is observed at t , $\tilde{z}_{it}$ is then observed. Thus, under MAR,

$$\Delta_{ixt} = \Pr(r_{ixt}=1 | z_i) = \Pr(r_{ixt}=1 | \tilde{z}_{it}).$$

We first model the one-step transition probability of the occurrence of missing data using logistic regression:

$$\text{logit}(p_{ixt}) = \text{logit}(E(r_{ixt}=1 | r_{ix(t-1)}=1, \tilde{z}_{it})) = \alpha_t + \beta_t^\top \tilde{z}_{it}. \quad (8)$$

By invoking MMDP and the assumption of no missing data at $t = 1$, we obtain:

$$\Delta_{ixt} = \Pr(r_{ixt}=1, r_{ix(t-1)}=1 | \tilde{z}_{it}) = p_{ixt} \Pr(r_{ix(t-1)}=1 | \tilde{z}_{i(t-1)}) = \prod_{s=2}^t p_{ixs}.$$

Now consider the case with $L = 1$. At each time t , the pair $z_{it} = (x_{it}, y_{it})^\top$ may be missing one or both components, yielding four possible missing data patterns indexed by a four-level nominal variable γ_{it} :

$$\gamma_{it}=1: \{r_{ixt}=1, r_{iyt}=1\}, r_{it}=2: \{r_{ixt}=0, r_{iyt}=1\},$$

$$\gamma_{it}=3: \{r_{ixt}=1, r_{iyt}=0\}, r_{it}=4: \{r_{ixt}=0, r_{iyt}=0\}.$$

A popular choice for modeling γ_{it} is the multinomial logistic model (e.g. McCullagh and Nelder, 1989):

$$p_{ilt}=E(\gamma_{it}=l | \gamma_{i(t-1)}=1, \tilde{z}_{it})=\frac{\xi_{ilt}}{\sum_{m=1}^4 \xi_{itm}}, 1 \leq l \leq 4, 2 \leq t \leq T, \quad (9)$$

$$\xi_{ilt}=\exp(\alpha_{it}+\beta_{it}^T \tilde{z}_{it}), \alpha_{i1}=0, \beta_{i1}=\mathbf{0}, \beta_{it}=(\beta_{it1}, \dots, \beta_{it(t-1)})^T,$$

$$\alpha_t=(\alpha_{t2}, \alpha_{t3}, \alpha_{t4})^T, \beta_t=(\beta_{t2}^T, \beta_{t3}^T, \beta_{t4}^T)^T, \zeta_t=(\alpha_t^T, \beta_t^T)^T, \zeta=(\zeta_2^T, \dots, \zeta_T^T)^T.$$

Under the assumption of no missing data at baseline ($t=1$), $p_{i1l}=1$ if $l=1$ and 0 if otherwise. Thus, we have:

$$\Delta_{ixyt}=\Pr[r_{ixt}r_{iyt}=1 | \tilde{z}_{it}]=p_{it1}\Pr[r_{ix(t-1)}r_{iy(t-1)}=1 | \tilde{z}_{i(t-1)}]=\prod_{s=2}^t p_{is1},$$

$$\Delta_{ixt}=(p_{it1}+p_{it3})\prod_{s=2}^{t-1} p_{is1}, \Delta_{iyt}=(p_{it1}+p_{it2})\prod_{s=2}^{t-1} p_{is1}.$$

In many applications, p_{ilt} may depend only on the most recently observed response at $t-1$, and the predictor in (9) under this Markov condition is simple, $\xi_{ilt}=\exp(\alpha_{it}+\beta_{it}^T z_{i(t-1)})$.

Note that for lag time $L=2$, the number of potential missing data patterns at each time t will increase at the rate of $(L+1)^2$. Although possible, it is likely to be difficult to model and estimate parameters given insufficient replications of missing data patterns in most applications. On the other hand, large lag times are unlikely to occur under MAR in real studies. For example, within the context of the HIV prevention example, it is unlikely that one response (diary or retrospective) will be continuously observed while the other is not. Thus, models with $L=1$ provide reasonable approximations for modeling missingness under MAR in most applications.

3 Simulation Study

3.1 Simulation Method

We conducted a limited simulation study to examine the performance of the proposed approach under the small and moderate sample sizes ($n=30, 65, 100$). For each sample size,

we generated pairs of continuous outcomes over three assessment times, $\mathbf{z}_{it} = (\tilde{x}_{it}, \tilde{y}_{it})^\top$, by simulating $\mathbf{z}_i = (\tilde{z}_{i1}^\top, \tilde{z}_{i2}^\top, \tilde{z}_{i3}^\top)^\top$ from a six-variate normal with mean $\mathbf{0}$ and variance $\Sigma = \Sigma_2 \otimes \Sigma_1$, where $\otimes$ denotes the Kronecker product of $\Sigma_2 = C_3(\rho_b)$ and $\Sigma_1 = C_2(\rho_w)$, Σ_1 the within-pair correlation matrix, $\rho_w = \text{Corr}(\tilde{x}_{it}, \tilde{y}_{it}) = 0.8$, Σ_2 the between-time correlation matrix, $\rho_b = \text{Corr}(\tilde{x}_{is}, \tilde{x}_{it}) = 0.5$, $s = 1, \dots, t$, $C_m(\rho)$ the $m \times m$ compound symmetry correlation matrix with correlation ρ . Next, we transformed each of the continuous outcomes $\tilde{x}_{it}$ and $\tilde{y}_{it}$ into a 4-level categorical response by:

$$x_{it} = 1I(\tilde{x}_{it} \leq \Phi^{-1}(0.3)) + 2I(\Phi^{-1}(0.3) < \tilde{x}_{it} \leq \Phi^{-1}(0.7)) + 3I(\Phi^{-1}(0.7) < \tilde{x}_{it} \leq \Phi^{-1}(0.9)) + 4I(\Phi^{-1}(0.9) < \tilde{x}_{it}), 1 \leq t \leq 3,$$

where Φ^{-1} denotes the inverse of the cumulative distribution function of the standard normal distribution. The true value of the Kendall's $\tau_b = (\tau_{b1}, \tau_{b2}, \tau_{b3})^\top$ is $\tau_{b1} = \tau_{b2} = \tau_{b3} = 0.643$ with τ_{bt} denoting the τ_b at time t .

For a given sample size n , we computed the estimate of τ_b and estimated its asymptotic variance Σ_{τ_b} using Theorem 1. The empirical type I error rate for testing $H_0: \tau_{b1} = \tau_{b2} = \tau_{b3}$

was obtained from $\hat{\delta} = \frac{1}{1000} \sum_{j=1}^{1000} I_{\{T_{nj}^2 \geq G_{0.95}\}}$, where T_{nj}^2 is the Hotelling's T-square statistic from the j th MC replication and $G_{0.95}$ is the 95th percentile of the central $F_{2, n-2}$ distribution.

For the missing data case, we assumed no missing data at baseline $t = 1$ and simulated the missing response according to a MCAR and MAR model. For MAR, we considered the MMDP for bivariate outcomes with lag time $L = 0$ and simulated the missing data indicators r_{it} at each time t from a Bernoulli distribution with the probability of success equal to p_{it} ($t = 2, 3$). We specified the one-step transition probabilities p_{it} at time t according to the logistic model in (8) under the Markov assumption, which in this case had the following form:

$$\text{logit}(p_{it}) = \beta_{t0} + \beta_{t1}(x_{i(t-1)} + y_{i(t-1)}), t = 2, 3.$$

We fixed $\beta_{t1} = 0.2$ ($t = 2, 3$) and solved the following equations for β_{20} and β_{30} to generate missing response rates of approximately 10% and 15% for times 2 and 3, respectively:

$$0.9n = \sum_{i=1}^n p_{i2}, 0.85n = \sum_{i=1}^n p_{i2}p_{i3}. \quad (10)$$

To ensure that the missing data indicators r_{it} followed the MMDP, we further imposed the following restrictions, $r_{i3} = r_{i2} \times r_{i2}$ ($1 \leq i \leq n$). For MCAR, the same approach above was used except for setting $\beta_{t1} = 0$ in (10).

In addition, power analysis for testing the null hypothesis of equal Kendall's τ_b s over time was conducted in a second simulation study within the same longitudinal data setting under complete data, MCAR, and MAR. The true value was set at $\tau_{b1} = 0.603$, $\tau_{b2} = 0.754$, and

$$\tau_{b3} = 0.809. \text{ The power was estimated by } \hat{\lambda} = \frac{1}{1000} \sum_{j=1}^{1000} I_{\{T_{nj}^2 \geq G_{0.95}\}}.$$

3.2 Simulation Results

Shown in Table 1 are the averaged $\hat{\tau}_b$, the averaged variance $\hat{\Sigma}_{\tau_b}$, along with the empirical variance $\text{Var}(\hat{\tau}_b)$ and type I error rate $\hat{\delta}$ over 1,000 Monte Carlo (MC) replications at each time point under different sample sizes. Under complete data, MCAR, and MAR, bias of $\hat{\tau}_b$ decreased and the empirical type I error rate showed a converging trend toward 0.05 as sample size increased. To examine the precision of the estimation, we compared the

variance $\hat{\Sigma}_{\tau_b}$ with the empirical variance $Var(\hat{\tau}_b)$. Overall, under complete data, it achieved the smallest bias and type I error, and the highest precision, while the largest bias and type I error occurred under MAR.

Shown in Table 2 are the averaged $\hat{\tau}_b$, the averaged variance $\hat{\Sigma}_{\tau_b}$, along with the empirical variance $Var(\hat{\tau}_b)$ and estimated power $\hat{\lambda}$ over 1,000 Monte Carlo (MC) replications. Similar patterns in power were found across complete data, MCAR, and MAR. Although the estimated power was low at $n = 30$; it increased rapidly as the sample size increased.

4 Application

We illustrate the method with data from a sexual health study. We set the statistical significance at $\alpha = 0.05$ for inference. A computational program for the analysis implemented in *R* is available upon request.

4.1 A longitudinal study in HIV prevention research

A longitudinal study in HIV prevention research was recently conducted to examine the accuracy of retrospective recall of sexual behavior using daily diary as a reference standard (Morrison-Beedy *et al.*, 2006). Although a daily diary may not be 100% accurate, this contemporaneous monitoring strategy addresses some of the key limitations of retrospective assessment such as recall bias (Schroder *et al.* 2003; Shrier *et al.* 2005). A sample of 54 adolescent girls monitored their sexual behavior with a daily diary and returned for assessment monthly for three months plus baseline. One of the primary interests is to assess recall bias due to assessment intervals and identify the degree of concordance between these two recall measurements using Kendalls τ_b at multiple time points. For illustration purposes, we focused on the reporting of unprotected vaginal sex. There was about 15% missing data in the outcome across the follow-up assessments.

Let $z_{it} = (x_{it}, y_{it})$ denote a four-level ordinal scale (no, less frequent, frequent, or highly frequent) of unprotected vaginal sex from monthly retrospective recall (x_{it}) and daily diary (y_{it}), $1 \leq t \leq 4$. Since missing data were the result of missed visits in this study, the MMDP with $L = 0$ applied and the missingness was modeled using the logistic regression in (8), with the predictor defined as a function of both observed x_{it} and y_{it} . As subjects were asked to recall their sexual behavior in the month prior to baseline ($t = 1$), the predictor of the logistic model at $t = 2$ (first monthly visit) contained only retrospective recall x_{i1} at baseline, rather than both retrospective and diary outcomes for the later visits at $t = 3, 4$. Given the low frequency of missing response, we again modeled the occurrence of missing data under the Markov condition with predictors, $\alpha_2 + \beta_{21}x_{i1}$ (for $t = 2$) and $\alpha_t + \beta_{t1}x_{it-1} + \beta_{t2}y_{it-1}$ (for $t = 3, 4$), for the logistic model in (8). Shown in Table 3 are the estimates of $\beta_t = (\beta_{t1}, \beta_{t2})$ ($\beta_2 = \beta_{21}$) and their standard errors and corresponding p-values. Although there is insufficient ground to reject MCAR, we performed the analyses under both MCAR and MAR.

Shown in Table 4 are the estimated τ_{bt} s and their standard errors over time computed under MCAR and MAR. Slightly larger estimates of τ_{bt} s and standard errors across the followup assessments were obtained under MCAR compared to those under MAR. In both data settings, the estimates seem to show a larger τ_b at the second assessment ($t = 3$) compared to the first and the last. To assess the significance of this trend, we tested a linear contrast of the form:

$$H_0: \tau_{bt} = \tau, \text{ for all time points } t, 2 \leq t \leq 4, \text{ versus}$$

$$H_a: \tau_{bs} \neq \tau_{bt}, \text{ for some } 2 \leq s < t \leq 4,$$

with a Hotelling's T-square statistic T_n^2 . Despite the variation in τ_b over the course of study, the results imply no evidence of degrading accuracy of retrospective reporting over time and monthly retrospective recall relative to daily diary can provide a reliable measure to record the frequency of unprotected sexual behaviors (p-value=0.61 under MCAR and 0.8 under MAR).

5 Discussion

We developed a new estimate of Kendall's τ in a form of U-statistic for longitudinal study data in the presence of missing data, and discussed non-parametric inference under both complete and missing data. In particular, we extended the inverse probability weighting (IPW) approach to the bivariate outcome in the current study to address the bias caused by missing data under MCAR and MAR. As none of the existing methods has addressed missing data in a longitudinal data setting for Kendall's τ , the proposed approach fills a critical gap in the extant literature. Modeling Kendall's τ under missing data is challenging since a large amount of missing data patterns can be generated from a bivariate response with repeated measurements. However, by introducing the definition of lag time (L) and considering the two special cases ($L = 0, 1$) under monotone missing data patterns, the proposed approach can handle missing data under MCAR and MAR in most applications.

In the multivariate case where the ordinal outcomes $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{iK})^\top$; $K \geq 3$, the proposed method can be easily extended. In particular, consider a longitudinal setting with K -variate outcome variable over T assessments. Let

$$\mathbf{z}_{it} = (z_{i1t}, z_{i2t}, \dots, z_{iKt})^\top, \mathbf{z}_i = (\mathbf{z}_{i1}^\top, \mathbf{z}_{i2}^\top, \dots, \mathbf{z}_{iT}^\top)^\top,$$

where $\mathbf{z}_{it}$ is a K -dimensional outcome of subject i at time t , $K \geq 3$, $i = 1, 2, \dots, n$; $t = 1, 2, \dots, T$. We assume $\mathbf{z}_i$, $i = 1, 2, \dots, n$, are independent samples but z_{ikb} , $k = 1, 2, \dots, K$, may be dependent. Let τ_{bmlt} denote the Kendall's τ_b of z_{imt} and z_{ilb} , $1 \leq m < l \leq K$. Also, let

$$\boldsymbol{\tau}_{bt} = (\tau_{b12t}, \tau_{b13t}, \dots, \tau_{b(K-1)Kt})_{K(K-1)/2 \times 1}^\top,$$

$$\boldsymbol{\tau}_b = (\boldsymbol{\tau}_{b1}^\top, \boldsymbol{\tau}_{b2}^\top, \dots, \boldsymbol{\tau}_{bT}^\top)^\top.$$

Hence, $\boldsymbol{\tau}_b$ consists of all pair-wise Kendall's τ_b among the K coordinates at every time point. By defining a vector $\boldsymbol{\eta}$ similar to (3), we obtain an unbiased, consistent, and asymptotic normal estimate of $\boldsymbol{\tau}_b$: $\hat{\boldsymbol{\tau}}_b = \mathbf{f}(\hat{\boldsymbol{\eta}})$. Based on the asymptotic distribution of $\hat{\boldsymbol{\tau}}_b$, we may test linear contrast of the form (7).

The extension to multivariate ordinal outcomes could benefit research involving high-dimensional data. For example, In a microarray analysis (e.g. Lobenhofer *et al.*, 2002), each array i has $(K-1)$ genes and the gene-expression level is characterized by its color intensity z_{imt} at time t , $m = 2; \dots, K$; $t = 1, 2, \dots, T$. The goal is to identify the association between gene-expression levels and subject's disease severity z_{i1t} . By using the proposed approach, we not only can estimate the Kendall's τ_b of disease severity with each gene-expression

level, but we also can test whether there is any statistical difference between these correlations at a fixed time point and assess how the correlations change over time.

In addition, missing data may also occur in the multivariate case which can become complicated with more missing data patterns compared to the bivariate case. Work is currently underway to generalize the proposed approach to address this issue as well as to develop double robust estimates within a longitudinal setting.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the following grant: Clinical Translational Science Center (UL1-RR024996). We sincerely thank the editor and two anonymous reviewers for their constructive critiques that help improve this manuscript immensely.

References

- Anfosso L, Efferth T, Albini A, Pfeffer U. Microarray expression profiles of angiogenesis-related genes predict tumor cell response to artemisinins. *The Pharmacogenomics Journal*. 2006; 6:269–278. [PubMed: 16432535]
- Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics*. 1994; 50:1088–1101. [PubMed: 7786990]
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 44:837–845. [PubMed: 3203132]
- Drake RE, Osher FC, Wallach MA. Alcohol use and abuse in schizophrenia. A prospective community study. *J Nerv Ment Dis*. 1989; 177:408–414. [PubMed: 2746194]
- Ferguson TS, Genest C, Hallin M. Kendalls' Tau for Serial Dependence. *The Canadian Journal of Statistics*. 2000; 28:587–604.
- Guo X, Pan W, Connett JE, Hannan PJ, French SA. Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Statistics in Medicine*. 2005; 24:3479–3495. [PubMed: 15977302]
- Hoeffding W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*. 1948; 19:293–325.
- Huntington TG, Hodgkins GA, Keim BD, Dudley RW. Changes in the Proportion of Precipitation Occurring as Snow in New England (1949–2000). *J. Climate*. 2004; 17:2626–2636.
- Kang M, Sen PK. Kendall's tau-type rank statistics in genome data. *Applications of Mathematics*. 2008; 53:207–221.
- Kendall M. A New Measure of Rank Correlation. *Biometrika*. 1938; 30:81–89.
- Kendall, M.; Gibbons, JD. Rank Correlation Methods, 5th edition. A Charles Griffin Title. London: Edward Arnold; 1990.
- King TS, Chinchilli VM, Carrasco JL. A repeated measures concordance correlation coefficient. *Statistics in Medicine*. 2007; 26:3095–3113. [PubMed: 17216594]
- Kocher SC, Gupta RP. Competitors of the Kendall-Tau Test for Testing Independence Against Positive Quadrant Dependence. *Biometrika*. 1987; 74:664–666.
- Lehmann EL. Consistency and unbiasedness of certain nonparametric tests. *Ann. Math. Statist*. 1951; 22:165.
- Lobenhofer EK, Bennett L, Cable PL, Li L, Bushel PR, Afshari CA. Regulation of DNA replication fork genes by 17 β -estradiol. *Molecular Endocrinology*. 2002; 16:1219–1229.
- Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. New York: Wiley; 1987.

- Ma Y, Tang W, Yu Q, Tu XM. Modeling Concordance Correlation Coefficient for Longitudinal Study Data. *Psychometrika*. 2010; 75:99–119.
- Ma Y, Tang W, Feng C, Tu XM. Inference for Kappas for Longitudinal Study Data: Applications to Sexual Health Research. *Biometrics*. 2008; 64:781–789. [PubMed: 18047535]
- Mankuta D, Goldner I, Knafo A. Intertwin Birth Weight Differences and Conduct Problems in Early Childhood. *Arch Pediatr Adolesc Med*. 2010; 164:457–461. [PubMed: 20439797]
- McCullagh, P.; Nelder, JA. *Generalized Linear Models*. 2nd ed.. London: Chapman and Hall; 1989.
- Morrison-Beedy D, Carey MP, Tu XM. Accuracy of audio computer-assisted self-interviewing (ACASI) and self-administered questionnaires for the assessment of sexual behavior. *AIDS Behav*. 2006; 10:541–552. [PubMed: 16721506]
- Paulus GG, Grasbon F, Walther H, Villoresi P, Nisoli M, Stagira S, Priori E, De Silvestri S. Absolute-phase phenomena in photoionization with few-cycle laser pulses. *Nature*. 2001; 414:182–184. [PubMed: 11700551]
- Pirttilä J, Uusitalo R. A ‘Leaky Bucket’ in the Real World: Estimating Inequality Aversion using Survey Data. *Economica*. 2010; 77:60–76.
- Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *JASA*. 1995; 90:106–121.
- Schroder KEE, Carey MP, Venable PA. Methodological challenges in research on sexual risk behavior: II. Accuracy of self-reports. *Annals of Behavioral Medicine*. 2003; 26:104–123. [PubMed: 14534028]
- Seber, GAF. *Multivariate Observations*. New York: Wiley; 1984.
- Shrier LA, Shih M, Beardslee WR. Comparison of momentary sampling with diary and retrospective self-report methods of measurement. *Pediatrics*. 2005; 115:573–581.
- Sipahi I, Debanne SM, Rowland DY, Simon DI, Fang JC. Angiotensin-receptor blockade and risk of cancer: meta-analysis of randomised controlled trials. *The Lancet Oncology*. 2010; 11:627–636. [PubMed: 20542468]
- Snow BAS. The third moment of Kendall’s tau in normal samples. *Biometrika*. 1962; 49:177–183.
- Wood JT. A variance stabilizing transformation for coefficients of concordance and for Spearman’s rho and Kendall’s tau. *Biometrika*. 1970; 57:619–627.
- Zhang H, Liu CT, Wang X. An association test for multiple traits based on the generalized Kendall’s tau. *JASA*. 2010; 105:473–481. [PubMed: 20711441]

Averaged $\hat{\tau}_{bt}$ and $\hat{\Sigma}_{\tau_{bt}}$ over 1000 MC replications along with estimated empirical variance of $\hat{\tau}_{bt}$, empirical type I error rate $\hat{\delta}$ under complete data, MCAR and MAR. True value of τ_b is $\tau_{b1} = \tau_{b2} = \tau_{b3} = 0.643$.

Table 1

Sample size	30			65			100		
Parameter estimate				Assessment time (t)					
	1	2	3	1	2	3	1	2	3
Complete data									
$\hat{\tau}_{bt}$	0.645	0.642	0.647	0.643	0.642	0.640	0.643	0.642	0.643
$Var(\hat{\tau}_{bt})$	0.010	0.010	0.010	0.004	0.004	0.004	0.003	0.003	0.003
$\hat{\Sigma}_{\tau_{bt}}$	0.009	0.009	0.009	0.004	0.004	0.004	0.003	0.003	0.003
$\hat{\delta}$	0.061			0.056				0.050	
MCAR									
$\hat{\tau}_{bt}$	0.645	0.646	0.651	0.643	0.643	0.648	0.643	0.640	0.644
$Var(\hat{\tau}_{bt})$	0.010	0.016	0.024	0.004	0.006	0.009	0.003	0.004	0.006
$\hat{\Sigma}_{\tau_{bt}}$	0.009	0.012	0.015	0.004	0.005	0.006	0.003	0.003	0.003
$\hat{\delta}$	0.065			0.059				0.053	
MAR									
$\hat{\tau}_{bt}$	0.645	0.648	0.653	0.643	0.643	0.646	0.643	0.641	0.645
$Var(\hat{\tau}_{bt})$	0.010	0.016	0.025	0.004	0.006	0.010	0.003	0.004	0.006
$\hat{\Sigma}_{\tau_{bt}}$	0.009	0.012	0.015	0.004	0.005	0.006	0.003	0.003	0.003
$\hat{\delta}$	0.068			0.061				0.056	

Table 2

Averaged $\hat{\tau}_{bt}$ and $\hat{\Sigma}_{\tau_{bt}}$ over 1000 MC replications along with estimated empirical variance of $\hat{\tau}_{bt}$ power $\hat{\lambda}$ under complete data, MCAR and MAR. True value of τ_b is $\tau_{b1} = 0.603$, $\tau_{b2} = 0.754$, and $\tau_{b3} = 0.809$.

Sample size		30			65			100		
		Assessment time (t)								
Parameter estimate		1	2	3	1	2	3	1	2	3
		Complete data								
$\hat{\tau}_{bt}$		0.605	0.755	0.809	0.604	0.752	0.809	0.602	0.750	0.808
$V_{ar}(\hat{\tau}_{bt})$		0.011	0.006	0.005	0.005	0.003	0.002	0.003	0.002	0.001
$\hat{\Sigma}_{\tau_{bt}}$		0.010	0.006	0.004	0.005	0.003	0.002	0.003	0.002	0.001
$\hat{\lambda}$			0.401			0.708			0.862	
		MCAR								
$\hat{\tau}_{bt}$		0.605	0.756	0.810	0.604	0.749	0.793	0.602	0.752	0.811
$V_{ar}(\hat{\tau}_{bt})$		0.011	0.015	0.026	0.005	0.005	0.009	0.003	0.004	0.007
$\hat{\Sigma}_{\tau_{bt}}$		0.010	0.010	0.012	0.005	0.003	0.004	0.003	0.002	0.003
$\hat{\lambda}$			0.372			0.662			0.837	
		MAR								
$\hat{\tau}_{bt}$		0.605	0.758	0.815	0.604	0.744	0.800	0.602	0.753	0.810
$V_{ar}(\hat{\tau}_{bt})$		0.011	0.015	0.026	0.005	0.006	0.010	0.003	0.005	0.007
$\hat{\Sigma}_{\tau_{bt}}$		0.010	0.009	0.010	0.005	0.003	0.004	0.003	0.002	0.003
$\hat{\lambda}$			0.365			0.654			0.831	

Table 3

Estimates of logistic regression for modeling occurrence of missing data at each monthly assessment for the HIV prevention study.

Assessment time	Predictors	Estimates	Standard errors	P-values
Month 1 ($t = 2$)	Retrospective (x_{11})	-0.07	0.61	0.91
Month 2 ($t = 3$)	Retrospective (x_{12})	0.60	0.90	0.50
	Daily diary (y_{12})	-0.42	0.72	0.56
Month 3 ($t = 4$)	Retrospective (x_{13})	0.91	1.11	0.41
	Daily diary (y_{13})	0.24	1.04	0.82

Estimates of Kendall's τ_b over time, test statistics and p-values for testing the null of equal τ_b over time for the HIV prevention study under MCAR and MAR.

Table 4

$\hat{\tau}_{bt}$ over time (standard errors)			$H_0 : \tau_{b2} = \tau_{b3} = \tau_{b4}$	
	Month 1 ($t = 2$)	Month 2 ($t = 3$)	Month 3 ($t = 4$)	
			test statistic T_n^2	p-value
MCAR	0.674(0.065)	0.759(0.07)	0.731(0.055)	0.499
MAR	0.673(0.064)	0.73(0.068)	0.719(0.053)	0.224
				0.61
				0.8